DOCUMENT RESUME

ED 159 214 TM 007 522

AUTHOR Berk, Ronald A.
TITLE A Consumers' Guide to Criterion-Referenced Test Item Statistics.
PUB DATE Mar 78
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978) ; Page 19 is marginally legible due to print size and quality

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Achievement Tests; Complexity Level; *Criterion Referenced Tests; Guides; *Item Analysis; Item Sampling; Research Design; *Statistical Analysis; Test Construction; *Test Items; Test Validity
IDENTIFIERS Item Discrimination (Tests); Item Homogeneity

ABSTRACT

Sixteen item statistics recommended for use in the development of criterion-referenced tests were evaluated. There were two major criteria: (1) practicability in terms of ease of computation and interpretation and (2) meaningfulness in the context of the development process. Most of the statistics were based on a comparison of performance changes (pretest-posttest) or differences (uninstructed-instructed) between criterion groups. Descriptions and critiques of the difficulty, discrimination, and homogeneity indices are presented in the form of a "consumers" guide." It was found that the relatively complex indices offered few if any advantages over the ones that were manually calculable and easily interpretable. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159214

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald A. Berk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A CONSUMERS' GUIDE TO CRITERION-REFERENCED TEST ITEM STATISTICS

Ronald A. Berk

The Johns Hopkins University

Paper presented at the annual meeting of the
National Council on Measurement in Education

Toronto, March 1978

Reproduced in U.S.A.

TM007 522

Abstract

Sixteen item statistics recommended for use in the development of criterion-referenced tests were evaluated. There were two major criteria: (1) practicability in terms of ease of computation and interpretation and (2) meaningfulness in the context of the development process. Most of the statistics were based on a comparison of performance changes (pretest-posttest) or differences (uninstructed-instructed) between criterion groups. Descriptions and critiques of the difficulty, discrimination, and homogeneity indices are presented in the form of a "consumers' guide" It was found that the relatively complex indices offered few if any advantages over the ones that were manually calculable and easily interpretable.

As part of the development of a criterion-referenced test, it is essential that the items undergo a review process to determine whether they function or "behave" consistent with the purposes for which they were constructed. These purposes can be summarized twofold: (1) to measure the instructional and behavioral objectives and (2) to differentiate among masters and nonmasters of a given objective. The extent to which the items are valid for these purposes depends upon the outcomes of two analyses. The first is concerned with the establishment of item-objective congruence. This represents the degree of relationship between each item and the objective it is intended to measure. The second is referred to as item analysis. This involves procedures designed to express the degree of relationship between the intent of each item and the responses of students to each item. The former is judgmental; the latter is both judgmental and statistical.

An extensive treatment of item-objective congruence can be found in the papers by Dahl (1971) and Rovinelli and Hambleton (in press). This paper focuses on the item analysis process and the numerous item statistics that have been devised to provide quantitative evidence of item validity. At least 19 different statistics can be identified. Their sheer volume indicates the topic has received considerable attention by psychometricians. Their lack of use, however, suggests that little attention has been devoted to the statistics by practitioners. This may be due to any of the following:

(1) Given the variety of statistics in the literature, it is difficult to determine which one is best for a particular application;

(2) Most of the statistics cannot be computed efficiently without a computer program;

(3) Proponents of "domain-referenced testing" stress that selecting items on the basis of statistical characteristics weakens the interpretability of the "domain score" (Millman, 1974; Millman & Popham, 1974).

In order to facilitate the proper use of item statistics by teachers and evaluators, the statistics have been compiled and evaluated in the form of a "consumers' guide".

Selection of Criterion Groups

The first consideration in the use of any of the 19 item statistics is the selection of the students to whom the items should be administered. The majority of the statistics are based on pretest-posttest repeated measurements from one group or two independent measurements from two different groups. Only four statistics are computed from one set of data gathered from one group (Brennan, 1972; Engel and LaDuca, 1977; Epstein, 1977; Pettie and Oosterhof, 1976).

The choice of which groups to use depends upon their availability and the purpose of the test. Since a criterion-referenced test is employed in most instances to identify masters and nonmasters of a set of objectives, the question a teacher needs to answer is: Who should be able to answer the items correctly and who should not? Two categories of students are of interest -- those who would be successful (masters) and those who would be unsuccessful (nonmasters) on the items. For the majority of instructional programs for which tests are developed, these groups will consist of students who have and students who have not been exposed to the instruction. It is likely that most if not all students prior to instruction will perform poorly on the items. If this level of performance is not obtained, either the items are failing to function as expected or the

objectives have been taught previously. When the latter appears to be the explanation, the justification for the instruction needs to be examined. In the case of the second group which is tested after the instruction has been completed, performance should be quite high. That is, most if not all of these students should answer the items correctly.

The two groups of students chosen for the item analysis are often referred to as criterion groups. Their selection is based on some criterion of current or future performance. This may be success in the instructional program in which the test will be used or success in a subsequent program, i.e., unit, course, grade level. The former often suggests summative decision-making at one point in time whereas the latter denotes placement decision-making within a predictive context. Placement in a program that contains sequentially arranged units each represented by an instructional objective presumes that a student's success in any single unit is prerequisite for success in the following unit. Therefore, students who have successfully completed the entire program can serve as one of the criterion groups for validating the placement test items. The same group may also be employed in the validation of items used for summative decisions about individual mastery at the conclusion of the program.

A clear definition of the criterion is essential. Two approaches for identifying criterion groups are described below. Both are applicable to criterion-referenced tests designed for placement and summative decision-making. The advantages and disadvantages of the approaches are noted.

Preinstruction-Postinstruction Measurements

The preinstruction-postinstruction method has been an integral part of mastery learning and diagnostic-prescriptive teaching for several years (Block, 1971, 1973; Carroll, 1963, 1970). It involves testing one group of students twice,

before (preinstruction) and after the instruction (postinstruction). Students are commonly tested with the same set of items on the two occasions. Alternatively, a parallel or equivalent set of items can be used on the second administration.

Advantage. The advantage of measuring students before and after instruction is that it permits the teacher to examine individual as well as group gains. This can be particularly useful in assessing progress from one unit to the next. Information on gains cannot be obtained from any other type of design.

Disadvantages. One major disadvantage of the approach, however, pertains to the impracticality of administering the posttest. It precludes the item analysis from being conducted before the instruction is completed. Hence, any instructional application of the measure would likely be deferred until the following year. A second disadvantage concerns the amount of time allowed between the pretest and posttest. When the administrations are in close succession, it is possible for the students' performance on the pretest to influence their performance on the posttest. One way to obviate this problem of testing effect would be to develop an equivalent set of items. If this is unfeasible due to practical constraints or the nature of the behavioral objectives (e.g., extreme specificity), it is recommended that the intervening period between the administrations be extended a reasonable length of time so that the carry-over effect of memory from the initial testing can be minimized. This extension should not be so great as four months or longer to avoid incurring the effect of maturation. It is well known that simply growing older can improve a student's achievement performance. This would confound the measurement of gain to which the teacher wishes to attribute the effect of instruction. Improvement as a result of instruction and maturation become inextricably mixed together when several months pass before the posttest is administered. The item analysis is designed to focus only on the change in item responses due to instruction.

Scoring. In order to expedite computation of the item statistics the results of the two measurements should be displayed in the form of an item X student matrix. This format is shown in Table 1. The table consists of the item scores (0,1) of 10 students who were administered five items both before and after instruction. The students were assigned a score of 1 for a correct response and 0 for an incorrect response or no response (omit).

The rationale for scoring an omit as 0 relates to the purpose of the test. Since a criterion-referenced test assesses how much a student knows unaffected by speed of response, no time limit is imposed. Therefore, when students are given ample time to complete the test, it can be assumed that if a student did not respond to an item it was because he/she did not know the correct answer.

Uninstructed-Instructed Groups

The uninstructed-instructed groups approach is a form of the known-groups technique employed for many years in the validation of personality tests (Cronbach and Meehl, 1955; Edwards, 1970; Wiggins, 1973). Only recently has the technique been applied to criterion-referenced measurement (Henrysson and Wedman, 1974; Jackson, 1970; Klein and Kosecoff, 1976; Millman, 1974; Zieky, 1972). It involves testing two separate groups of students at the same point in time, one group that has not received instruction (uninstructed group) and a second group that has received instruction (instructed group).

The known-groups technique compares one group of persons known by independent means to possess more of the specified trait/attribute with a second group known possess much less. In the validation of a criterion-referenced test, the trait of interest is knowledge. The acquisition of knowledge in most content domains is a continuous process. Learning to read, write, and solve mathematical problems occurs gradually throughout a person's formal schooling. Improvement in these areas is frequently accompanied by improvement in the other subject areas. Since students learn at different rates, at any given time the students within a single

Table 1
Item X Student Matrix for Preinstruction-Postinstruction Measurements

	Preinstruction					Postinstruction				
	Item					Item				
Student	1	2	3	4	5	1	2	3	4	5
1	0	0	1	1	1	1	0	1	1	1
2	0	1	1	1	0	1	1	1	1	1
3	0	1	0	1	1	1	1	1	1	1
4	0	0	0	1	1	1	1	0	1	1
5	0	0	1	1	0	1	0	0	1	1
6	0	0	0	1	0	1	1	0	1	0
7	0	1	1	1	1	1	0	0	1	1
8	0	0	1	1	0	1	0	1	1	0
9	0	0	0	1	0	1	1	0	1	1
10	0	0	1	1	0	1	1	1	1	1

Table 2
Item X Student Matrix for Uninstructed and Instructed Groups

	Uninstructed Group						Instructed Group				
	Item						Item				
Student	1	2	3	4	5	Student	1	2	3	4	5
1	0	0	1	1	1	1	1	1	1	1	1
2	0	1	1	1	0	2	1	1	0	1	1
3	0	1	0	1	1	3	1	1	1	1	1
4	0	0	0	1	1	4	0	0	1	1	1
5	0	0	1	1	0	5	1	0	1	1	0
6	0	0	0	1	0	6	1	1	1	1	1
7	0	1	1	1	1	7	1	0	0	1	1
8	0	0	1	1	0	8	0	1	0	1	0
9	0	0	0	1	0	9	1	0	1	1	0
10	0	0	1	1	0	10	1	1	1	1	1

classroom would vary in the amount of knowledge they possess. Therefore, in general, it can be expected that students in an instructed group will possess more knowledge related to a set of instructional objectives and an uninstructed group will possess much less. There are only two special cases where a group might know all or nothing. They will be discussed shortly.

The uninstructed-instructed groups approach requires that the two groups be identified on the basis of independent, informed teacher judgment. Students in one classroom who have not received instruction or who are known to be deficient on the objectives to be assessed may serve as the uninstructed group. Conversely, students in another classroom who have received "effective" instruction on the objectives may comprise the instructed group. This group, in the judgment of the teacher, should have demonstrated success in the instructional unit. Whether this success pertains to a current or subsequent unit depends on the orientation of the test. Student performance on written assignments, quizzes, and other classroom tasks is often the available information a teacher must rely upon for guidance in making the judgment.

The students chosen for the instructed group should be as similar as possible to those in the uninstructed group in ability level and all other relevant characteristics. Although the groups will often be unequal in size, the proportional distributions of the characteristics should be equivalent. For example, if one group has 60 percent boys and 40 percent girls, this balance should be maintained in the second group. The only real difference that should exist between the uninstructed and instructed groups is exposure to the instructional treatment.

While the approach is applicable in most testing situations, there are two instances where possessing or not possessing the knowledge necessary to master objectives will be characterized as all-or-nothing: (1) where minimum essential skills or competencies are being measured and (2) where no prior knowledge on objectives can be expected due to the nature of the content domain. The former is

concerned with the "all"; the latter emphasizes the "nothing".

Minimum essentials. When basic skills or minimum essentials are to be assessed, complete mastery by the instructed group is desirable. If a test is designed to identify children who cannot succeed or function at a minimum level in the "normal" instructional setting, the criterion successful group that is chosen must unequivocally demonstrate mastery or 100 percent performance. The criterion unsuccessful group should perform less than total mastery. This may range from 0 to less than 100 percent. The group should consist of the type of children to be identified by the test.

The item validation of a placement test to determine reading readiness, for example, would require the selection of children who clearly can and cannot read. In other programs such as learning disabilities and English as a second language the criterion groups would consist of children who possess and do not possess specific disabilities and who are native English speakers and speakers of English as a second language (e.g., Hispanic Americans), respectively. Evidence of these different attributes can often be found in school records. Again, teacher judgment plays a significant role in locating the children.

The procedure has implications for other competency-based decisions in course placement, grade level placement, high school graduation certification, and functional literacy. All extensions of the validation model must confront the same crucial issue - defining the criterion of success operationally in terms of the designated target groups.

No prior knowledge. Dissimilar to the above where 100 percent mastery is desirable for the instructed group, is the case where no prior knowledge can be expected for the uninstructed group. This occurs in courses that introduce the student to completely unfamiliar content, e.g., foreign language. A test administered prior to the instruction would yield results that unequivocally demonstrate nonmastery or 0 performance. For the purposes of validation there is no need to

administer the test to uninstructed students who are expected to know nothing about the content. Item and test performance can be assumed to equal 0. The test must be administered, however, to an instructed group.

Advantage. The advantage of the uninstructed-instructed group approach is one of practicality. The item analysis can be conducted at one point in time, prior to instruction, if an instructed group of students is concurrently available. This can facilitate the immediate use of the test for mastery-nonmastery decisions during the program being taught.

Disadvantages. One disadvantage is the difficulty of defining suitable criteria for identifying the groups. The more objectivity that can be incorporated into the process the better. "Informed" judgment as described previously is one step in this direction. At present, this seems to be an intuitively sound procedure for performing item analyses at the classroom level. Another disadvantage that is inherent in the approach is the problem of assuring group equivalence. Since randomly assigning students to instructed and uninstructed groups is generally unfeasible and inappropriate, the test-maker may need to explain differences in group performance that may not be attributable to the instructional treatment, e.g., age, sex, ability levels, socioeconomic background. Realistically, the best strategy to avert this charge from arising would be to deliberately control for possible extraneous factors in the initial selection of the groups and exhibit caution in interpreting the item analysis results.

Scoring. The scoring and presentation of the item data for the uninstructed and instructed groups is identical to the preinstruction-postinstruction approach. Table 2 displays an item X student matrix with item scores on five items administered to one group of 10 uninstructed students and a second group of 10 instructed students.

Item Statistics

The total of 19 statistics were evaluated according to two criteria: (1) practicability in terms of the ease with which the statistic could be computed and interpreted and (2) meaningfulness in the context of the test development process. Only five statistics appeared to satisfy both criteria. They include a difficulty index and four discrimination indices (Cox and Vargas, 1969; Klein and Kosecoff, 1976; Kosecoff and Klein, 1974; Roudabush, 1973). Among the 14 other statistics, seven were discrimination indices (Brennan, 1972, 1974; Darlington and Bishop, 1966; Haladyna, 1974; Kosecoff and Klein, 1974; Millman, 1974; Saupe, 1966) and four were measures of item homogeneity (Epstein, 1977; Engel and LaDuca, 1977; Pettie and Oosterhof, 1976; Popham, 1971). These were judged less practical than the previous indices since they were not manually calculable. In addition, their meaningfulness was less than clear. The remaining three statistics consisted of item stability (Sabers and Kania, 1971), item equivalence (Edmonston, Randall, and Oakland, 1971), and Bayesian indices (Estes and Estes, 1975). These seemed less useful than the preceding and, consequently, were not subject to further evaluation.

Item Difficulty

The difficulty of an item is the percentage of persons who answer the item correctly (percent pass). It is computed using the formula

$$DIFF = \frac{C}{N} \times 100,$$

where:

C = number of students who answer the item correctly

N = total number of students in the group

The index is equivalent to the mean of the item (proportion of students who got the item right) multiplied by 100.

Difficulty index values can range from 0 to 100. The higher the index, the easier the item. When all students answer the item correctly, the index equals 100; when all students respond incorrectly, the index is 0.

It is important that estimates of difficulty are obtained for both criterion groups. If difficulty levels are determined only for those students who have received the instruction, there is no way of knowing whether the students would have performed similarly without the instruction, e.g., an easy item after instruction may have also been easy prior to instruction.

Item Discrimination

The index of item discrimination measures the performance changes (pretest-posttest) or differences (uninstructed-instructed) between the criterion groups. This is consistent with the notion that a criterion-referenced test should maximize discriminations between groups and minimize discriminations among individuals within any one group (Glaser, 1963). Differences that do occur between the groups are presumed attributable to the effect of instruction. The interpretation of the index in terms of the instructional program has prompted some proponents of the technique to refer to it as a measure of instructional sensitivity.

Manually calculable indices. The vast array of indices found in the literature suggest that there is no single best method for estimating discrimination. Invoking the "law of parsimony" in this context, it is possible to extract four approaches that are conceptually and computationally simple yet statistically sound. They are evaluated in Table 3. The first two indices are based on the item difficulties expressed as proportions; the last two are based on the item response outcomes of the preinstruction-postinstruction measurements, i.e., incorrect-correct (0 → 1) and incorrect→incorrect (0 → 0).

Table 3

Evaluation of Item Discrimination Indices
(All were judged manually calculable and easily interpretable)

Discrimination Index	Statistic	Source	Comments
Pretest-posttest difference (DIS_{PPD})	Proportion	Cox and Vargas (1966); Tucker and Vargas (1974); Vargas (1969)	Proportion of students who answered item correctly on posttest minus proportion who answered it correctly on pretest. Range of values is from -1.00 to +1.00. Disadvantage: Index is not sensitive to individual performance changes, only total group gain or loss.
Uninstructed - instructed group difference (DIS_{UIGD})	Proportion	Klein and Kosecoff (1976); Levin and Marton (1971); Marton (1973)	Proportion of students in instructed group who answered item correctly minus proportion in uninstructed group who answered it correctly. Range of values is from -1.00 to +1.00. Correlates .70 and .12 with DIS_{PPD} (Crehan, 1974; Wedman, 1973). Relationship with DIS_{PPD} is contingent upon how closely instructed group resembles postinstruction grou Disadvantage: Same as DIS_{PPD}, but is a function of data bas rather than statistic.
Individual gain (DIS_{IG})	Proportion	Roudabush (1973)	Proportion of students who answered item incorrectly on pre test and correctly on posttest. A correction-for-guessing (Marks and Noll, 1967) is optional. Range of values is fro -1.00 to +1.00 with correction and from 0 to +1.00 without correction. Disadvantage: Assumptions underlying correctio for-guessing are questionable.
Net gain (DIS_{NG})	Proportion	Kosecoff and Klein (1974)	Extends DIS_{IG} by considering performance of all students wh answered item incorrectly on pretest; DIS_{IG} minus proportio of students who answered item incorrectly on both occasions Correction for guessing is optional. Range of values is from -1.00 to +1.00 with or without correction. Represents net individual gains/losses as a result of instruction. Yields more conservative values than preceding indices. Disadvantage: Same as DIS_{IG}.

The former are the simplest in that they are computed directly from the difficulty levels for the criterion groups; the latter are the most sensitive to pretest-posttest item score gains.

The choice of which of the four indices to use may be determined initially by the criterion groups available for the analysis. There is only one index for the case of uninstructed-instructed groups. Among the three indices appropriate for preinstruction-postinstruction measurements, DIS_{PPD} appears to be the simplest and DIS_{NG} the most sensitive. DIS_{NG} uses more information and possesses a greater range of possible values than DIS_{IG}. It is also statistically more sensitive than DIS_{PPD} since it takes into account the individual gains and losses resulting from the instruction rather than just the total group gain or loss. Overall, the DIS_{NG} index yields a more conservative estimate of item discrimination than the two alternatives.

The foregoing measures of item discrimination provide evidence of item validity. In general, the intent of the test items to discriminate between students who have and students who have not been exposed to instruction on a set of objectives is concurrently validated with group performance on the items. The external criterion was informed teacher judgment about the students' success/failure in a well-defined segment of instruction. This served as the basis for selecting the criterion groups. Given the persistent validation problem of obtaining a satisfactory criterion of success, teacher judgment should prove adequate for estimating validity at the classroom level. Certainly a more reliable, comprehensive alternative would be desirable. At present, however, such an alternative does not seem available.

Since the empirical item validation process is inherent in the test item analysis, its interpretation is subject to the conditions under which the item analysis is performed. The fact that each analysis is uniquely related to the

set of objectives, students' abilities and attributes, and instructional procedures characteristic of each classroom implies that test items that are valid for one class may not be valid for other classes. Teachers should not attempt to generalize the results beyond the context for which the items were originally designed.

Complex indices. Seven other discrimination indices were also identified. They are displayed in Table 4. These indices cannot be computed efficiently without a computer program. Hence, they may be of limited value and utility for teacher-made tests and assessments at the local level. Most of them do not simply measure the extent to which an item differentiates between criterion groups. Additional factors are considered, e.g., interitem correlations, that tend to complicate the computation and interpretation of item discrimination. For this reason the indices are listed in order of increasing overall complexity. The major disadvantages or deficiencies of the indices are noted.

Item Homogeneity

Four statistics that relate to item-objective congruence are evaluated in Table 5. These homogeneity measures are intended to statistically verify that items congruent with an instructional objective behave similarly on a single testing or on repeated testings (pretest-posttest). The assumption that the items should yield identical difficulty indices or change scores is questionable. Such "homogeneity" may be unrealistic and, in fact, undesirable in cases where a variety of skills may define the content of an instructional objective that encompasses several weeks of instruction or longer. The generality-specificity of the objective will probably determine the meaningfulness of the homogeneity statistic.

Table 4

Evaluation of Item Discrimination Indices
(Listed in order of increasing overall complexity)

Discrimination Index	Statistic	Source	Comments
Maximum possible gain	Proportion	Brennan (1974); Brennan and Stolurow (1971)	Extension of DIS_{PPD} index; proportion of maximum possible difference accounted for by obtained difference (DIS_{PPD}). Correlates .92 with DIS_{PPD} (Wedman, 1973). Disadvantages: Range of values and interpretations are different from most other discrimination indices.
B index	Proportion	Brennan (1972); Hsu (1971)	Criterion level is used to define master and nonmaster groups within one instructed group. Disadvantages: Validity of the criterion becomes a necessary condition for the validity of the item statistic. Index cannot be computed when all or none of the students meet the criterion. Interpretation of index values is rather unorthodox, i.e., ideal index is 0.
Internal sensitivity	Proportion	Koseoff and Klein (1974)	Criterion level is used to define master and nonmaster groups for both pretest and posttest. Disadvantage: Dependance on criterion.
Combined groups item-total r	Correlation	Kaladyna (1974); Helmstadter (1972)	Application of traditional item-total r to sample consisting of both criterion groups. Correlates from .64 to .86 with DIS_{PPD}. Index can be computed with most norm-referenced test item analysis programs such as ITEMAN (Bark and Griesemer, 1976). Disadvantage: Index is not sensitive to which students are scoring high and low (uninstructed or instructed) and, therefore, may not measure discrimination between designated criterion groups.
Item-criterion partial r	Correlation	Darlington and Bishop (1966); Millman (1974, pp.369-370)	Correlation of item scores with criterion group classification (0,1) holding total score constant. Several items per iteration method that takes into account interitem correlations is the recommended item selection procedure. Designed to maximize test validity. Disadvantages: Computation and item selection process are more complicated than that of preceding indices. Content and importance of objective being assessed are not considered. Index requires a specialized item analysis program or partial correlation program.
Change-item r	Correlation	Saupe (1966)	Correlation of item change scores (-1,0,+1) with total change scores from pretest to posttest. Designed originally to maximize reliability of change scores (Stanley, 1967; Webster, 1957) or validity with a criterion variable (Webster, 1956). Disadvantages: Most complex correlational approach requiring computation of six different covariances. Variance restriction on pretest or posttest scores will limit index values. Content and importance of objective being assessed are not considered in the item selection process. Index requires a specialized item analysis program such as CHANGE-SCORE (Bark and Griesemer, 1977).
_____	Multiple regression	Millman (1974, pp.367-368)	Stepwise regression analysis with items entered as predictor variables and criterion group classification serving as the criterion variable. Analysis incorporates interitem correlations. Disadvantages: Analysis requires a relatively large sample size. There is no item discrimination index; selection is based on inclusion in prediction equation. Item regression coefficients are used to weight item scores with no consideration given to the objective the items measure.

Table 5

Evaluation of Item Homogeneity Measures
(Listed in order of increasing overall complexity)

Statistic	Source	Comments
Chi-square	Pettie and Oosterhof (1976)	Modification of Popham's (1971) chi-square method for the case of one test administration (pretest only). Difficulty index of each item measuring an instructional objective is compared with the median difficulty of the set of items using chi-square. Disadvantages: Same as Popham's (1971) measure.
Chi-square	Popham (1971)	Pretest-posttest item frequencies are compared with median (expected) frequencies of a prototypic item using chi-square. Median values are derived from the frequencies of items measuring the same instructional objective. Items that behave differently from the prototype are considered for revision. Disadvantages: No critical value has been established for identifying abnormal items. Method is insensitive to the direction of the abnormality (Henrysson and Wedman, 1974).
_____	Engel and LaDuca (1977); Loevinger (1947)	Modification of Long's (1934) index of item homogeneity. Accounts for the relationship between item and total score. Conceptually similar to traditional index of item discrimination. Disadvantages: Interpretation of a zero index is unclear. Sampling distribution of statistic is unknown; therefore, no significance level can be determined. Item-total score phi coefficient has been the preferred approach (White and Saltz, 1957).
_____	Epstein (1977)	Average squared discrepancy index is designed to identify items that yield inconsistent response patterns. Based on the individual discrepancies of item scores from an item X student matrix. Interpretation of index values seems meaningful for criterion-referenced test items. Disadvantages: Computation is more complicated than preceding measures. No critical index value has been established for identifying "bad" items.

Conclusions

This "consumers' guide" was designed to synthesize the numerous and rather fragmented reports of criterion-referenced test item statistics into a context that is meaningful for persons charged with developing these tests, namely teachers and evaluators. The evaluations presented in Tables 3, 4, and 5 were conducted from the perspectives of practicability and the test development process. They suggest that the indices that are relatively complex offer few if any advantages over the ones that are manually calculable and easily interpretable. Probably the difficulty index and an appropriate discrimination index selected from Table 3 should provide sufficient information for estimating item validity in most applications. This should markedly improve the quality of criterion-referenced assessments at the classroom and district levels.

REFERENCES

Berk, R. A., & Griesemer, H. A. ITEMAN: An item analysis program for tests, questionnaires, and scales. Educational and Psychological Measurement, 1976, 36, 189-191.

Berk, R. A., & Griesemer, H. A. Change-score item and reliability analyses program. Applied Psychological Measurement, 1977, 1, 40.

Block, J. H. (Ed.) Mastery learning: Theory and practice. New York: Holt, Rinehart and Winston, 1971.

Block, J. H. (Ed.) Schools, society, and mastery learning. New York: Holt, Rinehart and Winston, 1974.

Brennan, R. L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Brennan, R. L. The evaluation of mastery test items (Final Report, Project No. 2B118). Washington, D.C.: National Center for Educational Research and Development, U. S. Office of Education, 1974.

Brennan, R. L., & Stolurow, L. M. An empirical decision process for formative evaluation. Paper presented at the meeting of the American Educational Research Association, New York, February 1971.

Carroll, J. B. A model of school learning. Teachers College Record, 1963, 64, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, 71-80.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, February 1966.

Crehan, K. D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.

Cronbach, L. H., & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Dahl, T. A. The measurement of congruence between learning objectives and test items. Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Darlington, R. B., & Bishop, C. H. Increasing test validity by considering inter-item correlations. Journal of Applied Psychology, 1966, 50, 322-330.

Edmonston, L. P., Randall, R. S., & Oakland, T. A model for estimating the reliability and validity of criterion-refernced measures. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1972.

Edwards, A. L. The measurement of personality traits by scales and inventories. New York: Holt, Rinehart and Winston, 1970. Chap. 4.

Engel, J., & LaDuca, A. Validity issues in the development of criterion-referenced tests. Paper presented at the meeting of the National Council on Measurement in Education, New York, April 1977.

Epstein, K. I. Predictive sample reuse and application for criterion-referenced test item analysis. Paper presented at the meeting of the American Educational Research Association, New York, April 1977.

Estes, C., & Estes, G. D. Criterion-referenced item analysis using Bayes' formula. Paper presented at the meeting of the National Council on Measurement in Education, Washington, D.C., April 1975.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Helmstadter, G. C. Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction. Paper presented at the meeting of the American Psychological Association, Honolulu, September 1972.

Henrysson, S., & Wedman, I. Some problems in construction and evaluation of criterion-referenced tests. Scandinavian Journal of Educational Research, 1974, 18, 1-12.

Hsu, T. C. Empirical data on criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, New York, February 1971.

Jackson, R. Developing criterion-referenced tests (TM Reports, No.1). Princeton, N.J.: Educational Testing Service, 1970.

Klein, S. P., & Kosecoff, J. B. Issues and procedures in the development of criterion-referenced tests. In W. A. Mehrens (Ed.), Readings in measurement and evaluation in education and psychology. New York: Holt, Rinehart and Winston, 1976. Pp. 276-293.

Kosecoff, J. B., & Klein, S. P. Instructional sensitivity statistics appropriate for objectives-based test items (CSE Report No. 91). Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Levin, L., & Marton, F. Provteori och provkonstruktion. Stockholm: Almqvist & Wiksell, 1971.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. Psychological Monograph, 1947, 61, No. 4 (Whole No. 285).

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Marton, F. Evalueringsteori och metodik. In G. Handal, L. G. Holmström, & O. B. Thomson (Eds.), Universitetsundervisning. Malmö: Studentlitteratur, 1973.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current practices. Berkeley, Calif.: McCutchan, 1974. Chap.6.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.

Pettie, A. A., & Oosterhof, A. C. Indices of item adequacy for individually administered mastery tests. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April 1976.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971. Pp. 79-98.

Roudabush, G. E. Item selection for criterion-referenced tests. Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research, in press.

Sabers, D. L., & Kania, J. G. Item precision in criterion-referenced measurement. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, April 1972.

Saupe, J. L. Selecting items to measure change. Journal of Educational Measurement, 1966, 3, 223-228.

Tucker, S. B., & Vargas, J. S. Item analysis of criterion-referenced tests for a large individualized course. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, April, 1974.

Vargas, J. S. Item selection techniques for norm-referenced and criterion-referenced tests. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Wedman, I. Reliability, validity and discrimination measures for criterion-referenced tests. Educational Reports, Umea, 1973, No. 4.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley, 1973. Chap.6.

Zieky, M. J. Methods of setting standards for criterion-referenced item sets. Princeton, N.J.: Educational Testing Service, 1973.